Sociology 400/500: Applied Social Science Statistics

Brandon M. Stewart Emily Cantrell Alejandro Schugurensky

Last Edited: August 10, 2020

Brandon M. Stewart
bms4@princeton.edu
brandonstewart.org

Emily Cantrell, Preceptor
emilymc@princeton.edu

Alejandro Schugurensky, Preceptor
as84@princeton.edu

This is the first class in Sociology's two-course graduate statistics sequence. Starting from only basic math, we build up a foundation for linear regression and its application to causal inference. The course is focused on the tools needed to do research and draws examples from across the social sciences. Students can take it both as a first course in linear regression or as a deeper dive into regression than a typical undergraduate sequence.

A Note on Timing:
The course is moving to primarily asynchronous content with one all-class course meeting of 1.5 hours per week. Optional synchronous office hours will be available by appointments which will be scattered from Tuesday to Friday. The course meetings will be split between undergraduates and graduate students. The times are:

- Undergraduates: Tuesdays 7:30pm-8:50pm Eastern
- Graduates: Wednesday: 10am-11:30am Eastern

1 The Basics

1.1 Course Goals

This is the first course in the Department of Sociology's two-course graduate statistics sequence. Starting from only basic math, we build up a foundation for linear regression and its application to causal inference. Students will learn the statistical and computational principles necessary to perform modern, flexible, and creative analysis of quantitative social data.

By the end of this semester, you will be able to:

- Critically read and reason about quantitative social science using linear regression techniques.
- Conduct, interpret, and communicate results from analysis using multiple regression.
- Explain the limitations of observational data for making causal claims and distinguish between identification and estimation in causal inference.
- Understand the logic and assumptions of several modern designs for making causal claims.
- Write clean, reusable, and reliable `R` code in the tidyverse style.
- Feel empowered working with data.

The second course in the course in the sequence, SOC 504, will be offered in the spring. The overarching goal of the two-course sequence is to move you from being consumers of quantitative research to producers of it. The capstone of the two-course sequence is the replication and extension project. In this project, completed in Sociology 504, you and a partner will choose a paper of interest, reproduce the results and then extend them to make something new. The projects are presented at Graduate Research Day in the Spring during a poster session and written up in a paper.

Upon finishing the two course sequence, you should be able to read an original scholarly article describing a new statistical technique, implement it in computer code, estimate the model with relevant data, interpret the results, and explain the results to someone unfamiliar with statistics. Beyond the two-course sequence, we encourage you to participate in the broader statistical life at Princeton including the Sociology Statistics Reading Group (`https://scholar.princeton.edu/bstewart/sociology-statistics-reading-group`), the Quantitative Social Science Colloquium (`https://qaps.princeton.edu/colloquium`) and the Center for Statistics and Machine Learning (`https://csml.princeton.edu/`).

This course will require a lot of hard work from all of us; however, we have structured the class to provide you the maximal return on every hour of work you put in. As you read through this syllabus you will find numerous avenues for seeking help. If you are willing to put in the time, we are always happy to help. Please don't be shy about telling us where you need support.

More broadly, the statistics sequence is designed to get you to a point where you can teach yourself new statistical methods by reading the literature. We can't teach you all the statistical techniques that you will need during your career, but we can prepare you to teach yourself.

1.2 Prerequisites

The most important prerequisite is a willingness to work hard on possibly unfamiliar material. Learning statistical methods is like learning a new language, and it will take time and dedication to master its vocabulary, its grammar, and its idioms. However like studying languages, statistics and programming yield to daily practice and consistent effort.

We intentionally have no formal pre-requisites. Beyond high-school level algebra, it is helpful to have some familiarity with univariate calculus (essentially knowing what derivatives and integrals are in principle even if you forget how to do the mechanics) and basic matrix operations (matrix multiplications and inverses). It will also be helpful if you have some experience with computer programming, specifically the R language. If these concepts are unfamiliar, you should review the

materials from our Department of Sociology summer methods camp (`http://pusocmethodscamp.org/`).

Even if you have seen some of the materials in class before (e.g., you had an undergraduate class on linear regression), you will likely find a lot to learn here. By rebuilding the foundations of linear regression from scratch, we help to ensure that everyone is on the same page, can more deeply appreciate the intricacies of these methods, and can have a solid foundation for learning more advanced methods. If you are concerned that you may have already covered the material before, come talk to the instructor.

1.3 Course Structure in the Time of COVID

This year poses exceptional challenges for the traditional structure of the course. Our plan is to move towards predominantly asynchronous content delivery with one hour each week to come together (digitally) and ask and answer questions. This means the components of the class break down as following each week.

- Lectures (Asynchronous):
  The lectures cover the core conceptual material of the course. Prerecorded video lectures will be posted each week. These will be broken into topic-specific chunks for ease of viewing but will have an intended viewing order with continuity from video to video. As an upper estimate these will total 150-160 minutes of lectures each week (the approximate length of scheduled class time).

- Precepts (Asynchronous):
  The precepts cover the mechanics of how to do the coding necessary to complete the problem sets. These videos will also be broken into topic-specific chunks and will likely total considerably less than the 110 minutes currently allocated for them in the schedule.

- Course Meeting (Synchronous):
  This will be an opportunity for you to ask me questions about the lecture and for us to review the material together. This will be the only mandatory synchronous course component.

- Office Hours (Synchronous, Optional)
  The instructor and the preceptors will offer extended office hours by sign up. Office hours with the instructor are primarily for discussing conceptual material while office hours with the preceptors should be used for assistance with the programming and/or problem sets.

We may need to make changes to this structure on the fly to meet student needs. Please keep in touch with us about what is and isn't working for you. It might take a few adjustments to get this right!

2 Materials

2.1 Computational Tools

The best way, and often the only way, to learn about data analysis and new statistical procedures is by doing. We will therefore make extensive use of a flexible (open-source and free) statistical

software program called R, RStudio, and a number of companion packages in the tidyverse style. Problem sets and the final exam will be completed in R Markdown. You will learn how to program in this class, if you do not know already.

2.2 Readings

This class uses extremely detailed lecture slides (it is not uncommon to have 100 slides in a week). We encourage you to think of these (and the lectures) as the primary reading in the class (although we might assign an occasional supplement here or there). If you are someone who benefits a lot from reading material and you find the slides aren't working for you, come talk to me and I will help find a reading that is well situated to your background and interests.

Sometimes though it is just helpful to have a reference book. We include a few by topic below.

- Programming
  - Grolemund and Wickham. 2017. *R for Data Science* (available online for everyone)
  - Healy, Kieran. 2018. *Data Visualization: A Practical Introduction.* Princeton University Press. (available online for everyone)
- Probability and Random Variables
  - Blitzstein and Hwang. 2014. *Introduction to Probability* (available online through the library)
- Regression
  - Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press. (available online through JSTOR)
  - Aronow and Miller. 2019. *Foundations of Agnostic Statistics.* (available online through the library)
  - Fox, John. 2016 *Applied Regression Analysis and Generalized Linear Models. 3rd Edition.*
  - Shalizi, Cosma. Forthcoming. *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press (Note that this book is still being written and you can find draft PDFs on the linked page above.)
- Causal Inference
  - Hernán, Miguel A. and James M. Robins. Forthcoming. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC. (Note that this book is still being written and you can find draft PDFs on the linked page above.)
  - Morgan, Stephen L, and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Second Edition.* Cambridge University Press. (available online through the library)

3 Assignments

There are three main types of assignments (summary here, details below):

1. **Preparing for the Course Meeting:** Before each course meeting, we expect you to have viewed the week's lecture videos and annotated any sources of confusion. These will be available in the online annotation system Perusall.

2. **Weekly problem sets:** Learning data analysis takes practice. The problem sets are described below.
3. **Final exam:** A cumulative take-home final exam will conclude the semester.

3.1 Preparing for the Course Meeting

Let's be honest, watching online lecture videos is hard. There is going to be a temptation to only pay half attention because you are folding laundry. A key part of this class is going to be resisting that temptation and watching carefully to prepare for class.

Normally in a lecture, if you didn't understand something I said, I'd ask you to interrupt me with a question so I can pause and re-explain. With prerecorded lectures you can pause-rewind-rewatch (a huge advantage over standard lectures!) but I can't dynamically respond to your question. This is why we are going to use the Perusall—an annotation platform that can handle video. As you encounter things you have a question or comment about, add an annotation in Perusall. Someone may be able to answer your question right there.

The annotations will also play a key role in helping to inform the topics we cover in the course meeting. In these course meetings I'm going to assume that everyone has already watched the lectures carefully—this is a time for consolidating learning, not seeing it for the first time! To that end, I really recommend sitting down and taking notes on the lectures, posting annotations on Perusall and engaging deeply with the material.

3.2 Problem Sets

Statistical methods are tools and it isn't very instructive to read a lot about hammers or watch someone else wield a hammer. You need to get your hands on a hammer or two. Thus, in this course, you will have homework on a weekly basis. The assignments will be a mix of analytic problems, computer simulations, and data analysis.

Format Assignments should be completed in R Markdown which allows you to show both your answers and the code you used to arrive at them. If you haven't seen R Markdown before, have a look at the materials on `http://pusocmethodscamp.org/`. Your wonderful preceptors will provide you with more detailed instructions before the first assignment is due.

Assessment Problem sets will be graded from 0-50 points. We also reserve the right to add bonus points for aesthetics including presentable graphs, clear code, nice formatting and well written answers. Solutions will be available directly after the problem set deadlines. The problem sets including looking at the solutions key is an extremely important part of the learning process, so please keep up with the work!

Extensions You can have *one* no questions asked extension of one week on a problem set of your choosing. If you don't take an extension, we will drop your lowest grade (of any partially completed problem set). If all your problem sets are completed and with top-level grades (such that dropping the lowest wouldn't help you), we will add a comparable grade bonus to your final exam. When submitting the work on which you claim the extension please include a note indicating the original date and that you are claiming your one extension; you do not need to explain why you are taking the extension. Because we do not want to hold up the class we will not wait for everyone to submit

their problem sets in order to post the solutions key. If you are turning your problem in late you are on your honor to *not look at the solutions* before submitting your work and you are required to explicitly write on your assignment that you have not looked at the solutions. If you exceed the one-week extension period your grade will drop on the problem set by 10% per day down to 30%. You have until the beginning of reading period to submit a late problem set to get the 30% minimum.

Extreme Circumstances In the time of COVID, I realize that the above extension policy may not be enough to deal with all situations. For example, if a student contracts COVID this might disrupt weeks of school work. If an extreme disruption occurs please reach out to me as quickly as possible. We will assess your situation and then in collaboration with your Director of Graduate Studies or Undergraduate Dean we will come up with a plan to get you up to speed as quickly as possible.

Collaboration Policy: Unless otherwise stated, we encourage students to work together on the assignments, but you should write your own solutions (this includes code). That is, *no copy-and-paste* from other people's code. You would not copy-and-paste from someone's paper, and you should treat code the same way. However, we strongly suggest that you make a solo effort at all the problems before consulting others. We know this is more complicated in the time of COVID and we will help you find people to work with.

3.3 Final Exam

The final assessment of the class is a take-home exam. The exam is "open-book" in the sense that you can use the slides, your notes, books, and internet resources to answer the questions. However, the final exam must be completed *by yourself.* The exam will be available during the entire period allocated for take-home exams by the university. It will be approximately the length of a long problem set (although we caution that it might take longer if you are used to collaborating on the problem sets). We encourage you to start early. Before the final exam we will distribute a practice final that will help you in your preparations.

3.4 Grading

Final grades will be a weighted average of the final exam (20%), class participation including Perusall annotations/Piazza posting (20%) and the weekly problem sets (60%).

4 How to Learn in this Course

If you find this course challenging, you are not alone. Statistics can be challenging and we cover a lot of ground. However, I am confident that you can handle it. In this section of the syllabus I'm going to provide details on some of the forms of support that we offer in this class and pull back the curtain a bit on the pedagogical design.

Your primary responsibilities in this class are to *work hard* and *communicate* with us about what you need. You can't learn if you aren't putting in the time. We can't help if we don't know there is a problem.

The course is designed to provide every tool we can think of to help you learn the material. If you are willing to put in the time, we want to ensure that time is used as effectively as possible.

4.1 Resources for Getting Help

There are a few main sources of support in the class.

1. Lectures, Slides and Precept
   Lectures and precepts are recorded which will help you if you want to revisit material. The slides themselves are also a great resource and will be available in PDF form. Be sure to go through them!

2. Perusall
   Annotate the videos on Perusall and use the platform to ask questions about material you don't understand. If you have a question, someone else probably does too. Do them a favor and post it first!

3. Course Meetings
   These sessions are going to be a bit of an experiment but the entire purpose is to help you fully understand the material. Participate! Ask Questions! Be in the moment!

4. Course Meeting Feedback
   When I teach in person, each lecture I pass around a note card and ask each student to write down something about class. They can write something they liked or didn't like. Something they want to understand better or want to hear more about it. Maybe they want to know how a piece of material connects to the broader goals of the class. They can even just draw a smiley face. I find this feedback to be invaluable because open lines of communication are key. We are going to do some digital form of the notecards.

5. Piazza
   Piazza is a classroom discussion board where you can post questions about the material. You will not be required to post, but the system is designed to get you help quickly and efficiently from classmates, the preceptors, and me. **Unless the question is of a personal nature or completely specific to you, you should not e-mail teaching staff**; instead, you should post your questions on Piazza. The course staff will be monitoring the page, but we encourage you to help your classmates as well. A big part of why we use Piazza is because reading other people's questions can be really helpful for bolstering your understanding of the material.

6. Office Hours
   Each of the preceptors will be offering multiple office hours this week via sign ups. These are primarily for questions about the week's problem sets. I will offer sign ups for office hours as well. These are primarily for broader conceptual questions.

7. Problem Set Solutions
   As soon as the problem sets are due the solution is posted. I know it is really tempting to just turn your focus to the next problem set, but if you were at all unsure about something, I highly encourage you to check the solutions.

8. Final Exam Prep
   We will host a review session for the final exam although we aren't yet sure what form it will take.

9. Individual Tutoring
   In circumstances where it is deemed necessary, the department has agreed to pay for individual and/or small group tutoring. This would be aimed at helping with basic programming in R and foundational concepts in the course. If you believe this would be beneficial to you, please contact the instructor.

If you can think of something else that would be useful to you, we encourage you to come talk to us. Again, if you are willing to put in the time, we can get you a form of support that matches your needs.

4.2 How is the Course Designed?

At a high level, this course builds up the infrastructure of linear regression and causal inference from the basics of probability. The first four weeks focus on foundation elements: probability, random variables and the basics of statistical inference. The second four weeks covers linear regression and its variants. The final four weeks are devoted to causal identification and estimation.

Each week covers a specific topic with a series of closely connected lectures included in a common slide deck. These lectures are designed to fill you in on the core statistical ideas animating the week's topics. The lecture won't focus on code, it will focus on the underlying logic and the applications of the ideas to social science research. The synchronous course meeting will be an opportunity to review tricky parts of the lecture and consolidate understanding.

Precept videos will teach you the programming tools necessary to implement the things shown in lecture. The code shown in precept is very closely tied to what you will need to complete the problem sets.

The problem sets are where I expect the majority of the learning will be solidified. These assignments are challenging and time consuming, but it is only through carefully engaging with the material that you will cement your understanding of it. If at the end of lecture you feel like you don't have a good handle on the material—that's to be expected. If after the problem set has been submitted you still feel uneasy with the material, you should come to office hours and talk to one of us about it.

Finally, there is a strong focus on the class on understanding why things work rather than just applying them. For example, we will often program up our own functions for things `R` has built in functionality for. Why do we do that? By programming it ourselves or deriving a known result, we force ourselves to really understand the underlying mechanics. This not only improves our understanding of statistical analysis but it also helps learn new things in the future. Our goal isn't just for you to learn this material, it is prepare you to teach yourself new material in the future.

4.3 Advice from Prior Generations of Students

Each year I ask students to provide advice to future generations of students. Here is some advice from prior students responding to "What advice would you give to another student considering taking this course?" I think the advice is great and it may be helpful coming from other students.

- Be ready to spend a lot of time
- Ask questions if you don't know what's going on!
- Study hard, work hard, review the slides.
- Investing a considerable amount of time in getting familiar with R and its various tools will pay off in the long run!
- Go over the lecture slides each week. This can be hard when you feel like you're treading water and just staying afloat, but I wish I had done this regularly.
- It's challenging but very doable and rewarding if you put the time in. There are plenty of resources to take advantage of for help.
- It will be hard but you will learn so much.
- This course is very challenging but greatly contributed to my understanding of social statistics. If you're truly invested in the subject and willing to put in the work (more than you expect possibly), it will be one of the best courses you've taken.
- This is a course where you will learn a lot and spend most of your time doing the psets. I highly recommend office hours for clarification as lecture covers a lot of material.

4.4 A note to everyone that is not a first year PhD student in the Department of Sociology

This course is completely designed to provide training to first year PhD students in the Department of Sociology. That means that some of the topics covered—and the way that these topics are covered—may not be optimal for your particular backgrounds. Undergraduates may find that the course has a different style and pace than the courses they have taken in the past. Those warnings aside, previous generations of students from many different departments have found the course useful.

5 Course Outline

The Rhythm of the Week

The working plan is to have everyone view the lectures/precepts for the week (adding annotations in Perusall) by the end of the day Monday. Tuesday night or Wednesday morning will be the course meeting. Work on the problem set is intended to go Tuesday to Thursday during which office hours will be available for sign up (problem set will be due Friday at 5PM eastern). You should then start watching the next week's lectures on Friday.

Two weeks below marked with * are half weeks and will not have a problem set at the end.

Week 1: Introduction and Probability (August 31)

- Course Details, Outline and Requirements
- Probability

NB: This week has only half the material of a usual week so it should be a gentler start.

Week 2: Random Variables (September 7)

- Random Variables
- Marginal, joint, and conditional distributions
- Expectations, Conditional Expectations
- Covariance, correlation, and independence

Week 3: Learning from Random Samples (September 14)

- Populations, samples, estimation
- Point estimation
- Properties of estimators
- Interval Estimation

Week 4: Testing and Regression (September 21)

- Hypothesis testing
- Nonparametric regression
- Parametric models and linear regression
- Bias-variance tradeoff
- Regression as a predictive model

Week 5: Simple Linear Regression (September 28)

- Mechanics of Ordinary Least Squares
- Assumptions of the linear model
- Properties of least squares
- Inference with regression

Week 6: Linear Regression with Two Regressors (October 5)

- Mechanics of regression with two regressors
- Simpson's Paradox
- Omitted variables and multicollinearity
- Dummy variables, interactions, and polynomials

Week 7* : Multiple Linear Regression (October 12, Fall Break)

- Matrix algebra and mechanics of multiple linear regression
- Classical Inference in a multiple linear regression model

NB: Fall break is at the beginning of this week but we will move the course meeting to later in the week but there will not be a problem set so it should still be a substantially lighter week.

Week 8: Rethinking Regression (October 19)

- Agnostic Regression
- Bootstrap
- Behavior of Regression

Week 9: Regression in the Social Sciences (October 26)

- Making Claims
- Visualization
- Frameworks for Causal Inference (Potential Outcomes and Causal Graphs)

Reading TBD

Week 10: Causality With Measured Confounding (November 2)

- The Assumption of No Unmeasured Confounding
- Choosing Conditioning Variables

Week 11: Unmeasured Confounding and Instrumental Variables (November 9)

- Natural Experiments
- Instrumental Variables
- Regression Discontinuity

Week 12: Repeated Observations and Panel Data (November 16)

- Fixed effects
- Difference-in-Differences
- Dynamic Models

Week 13*: Review and Final Discussion (November 23, Thanksgiving)

- Overview of the course
- Estimands

Reading: TBD

6 Inspirations

The development of that course was in turn influenced by a number of people particularly: Matt Blackwell, Dalton Conley, Adam Glynn, Justin Grimmer, Jens Hainmueller, Erin Hartman, Chad Hazlett, Gary King, Kosuke Imai, Kevin Quinn, Matt Salganik, and Teppei Yamamoto. I am grateful to everyone who has contributed to these materials, directly or indirectly. I am also grateful to generations of past preceptors who have had a huge influence on the direction the class has gone including Clark Bernier, Elisha Cohen, Alex Kindel, Ian Lundberg, Shay O'Brien, Ziyao Tian, and Simone Zhang.